SD-TSIA 211
Optimization for Machine Learning
29 January 2018

Paper documents are allowed (lecture notes, exercises and books)
Electronic devices are forbidden

Exercise 1 (Dual of the logistic regression problem)**.**
We consider the following logistic regression problem with ridge regularization given by

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{m} \log\big(1+\exp(-y_i a_i^\top x)\big) + \frac{\lambda}{2}\|x\|^2 \tag{1}$$

where for all $i \in \{1,\dots,m\}$, $y_i \in \{-1,1\}$ and $a_i \in \mathbb{R}^n$.

1. Let h such that $\forall u \in \mathbb{R}$, $h(u) = \log(1+\exp(u))$. Show that for all u, $h'(u) \in [0,1]$.
2. Show that h is convex.

We define the Fenchel transform of h by

$$\forall \alpha \in \mathbb{R}, h^*(\alpha) = \sup_{u\in\mathbb{R}} \alpha u - h(u)\,.$$

3. Let $\alpha \in]0,1[$. Show that

$$h^*(\alpha) = C_1(\alpha)\log(\alpha) + C_2(\alpha)\log(1-\alpha)$$

where $C_1(\alpha)$ and $C_2(\alpha)$ should be explicited.

4. Let $\alpha \notin [0,1]$. Show that $h^*(\alpha) = +\infty$.
5. Show that $h^*(0) = h^*(1) = 0$.
6. Give a convex function f and a matrix M such that for all $x \in \mathbb{R}^n$,

$$\sum_{i=1}^{m} \log\big(1+\exp(-y_i a_i^\top x)\big) = f(Mx)$$

7. By introducing a new variable $z \in \mathbb{R}^m$ and a linear constraint, define a problem equivalent to (1) of the form

$$\min_{x\in\mathbb{R}^n, z\in\mathbb{R}^m} f(z) + g(x) \tag{2}$$
$$\text{st} : Mx = z$$

Please explicit the functions f and g and the matrix M.

8. Write the Lagrangian associated to Problem (2).
9. Calculate the dual function.

Exercise 2 (Subgradient method)**.**
Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. We assume that f has a minimizer x^*.
The subgradient method is the following algorithm that starts at $x^0 \in \mathbb{R}^n$ and for all $k \in \mathbb{N}$:

$$\begin{aligned} &\text{Take } g_k \in \partial f(x_k) \\ &x_{k+1} = x_k - \gamma_k g_k \ . \end{aligned}$$

The sequence $(\gamma_k)_k$ is such that $\gamma_l > 0$ for all l, $\displaystyle\sum_{k=0}^{+\infty} \gamma_k = +\infty$ and $\displaystyle\lim_{N\to\infty} \frac{\sum_{k=0}^N \gamma_k^2}{\sum_{k=0}^N \gamma_k} = 0$.

1. Using the definition of the subgradient, show that for all $x \in \mathbb{R}^n$ and $g \in \partial f(x)$, we have
$$f(x+g) \geq f(x) + \langle g, g\rangle$$
2. Suppose that f is L-Lipschitz ($|f(x) - f(y)| \leq L\,\|x-y\|$). Show that there exists $M > 0$ such that for all $x \in \mathbb{R}^n$ and for all $g \in \partial f(x)$, we have $\|g\|_2 \leq M$.
3. Conversely, suppose that there exists $M \geq 0$ such that for all $x \in \mathbb{R}^n$ and for all $g \in \partial f(x)$, $\|g\|_2 \leq M$. Show that f is Lipschitz continuous.

For the rest of the exercise, we assume that f is L-Lipschitz.

4. Find $\beta(k)$ such that for all $k \in \mathbb{N}$,
$$\frac{1}{2}\|x_{k+1} - x^*\|_2^2 = \frac{1}{2}\|x_k - x^*\|_2^2 + \gamma_k\langle g_k, x_* - x_k\rangle + \beta(k)\,\|g_k\|_2^2 \ .$$
5. Show that for all $k \in \mathbb{N}$, $f(x^*) \geq f(x_k) + \langle g_k, x_* - x_k\rangle$
6. Deduce from this inequality that
$$\gamma_k\big(f(x_k) - f(x_*)\big) \leq \frac{1}{2}\|x_k - x^*\|_2^2 - \frac{1}{2}\|x_{k+1} - x^*\|_2^2 + \beta(k)M^2$$
7. Show that
$$\sum_{l=0}^{k} \gamma_l\big(f(x_l) - f(x_*)\big) \leq \frac{1}{\sum_{l=0}^k \gamma_l}\Big(\frac{1}{2}\|x_0 - x_*\|_2^2 + \sum_{l=0}^{k}\beta(l)M^2\Big)$$
8. Denote $\bar{x}_k = \dfrac{1}{\sum_{l=0}^k \gamma_l}\displaystyle\sum_{j=0}^{k}\gamma_j x_j$. Using the fact that $\bar{x}_k$ is a convex combination of the previous iterates, find a bound on $f(\bar{x}_k) - f(x_*)$.
9. Using the properties of the sequence (γ_k), show that $f(\bar{x}_k)$ converges to $f(x_*)$.

Exercise 3 (ℓ_1 regression).

In this exercise we study the following ℓ_1 regression problem

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_1 \tag{3}$$

where $A = \begin{bmatrix} a_1^\top \\ \vdots \\ a_m^\top \end{bmatrix}$ is a $m \times n$ matrix and $b \in \mathbb{R}^m$.

As the objective is affine by parts and lower bounded, it has at least one minimizer x_* (you do not have to show this fact).

The objective function is affine by parts, so it could be dealt with by linear optimization. However, we are going to show that algorithms based on subgradients can also solve this problem.

1. We define $f : \mathbb{R}^n \to \mathbb{R}$ such that $f(x) = \|Ax - b\|_1 = \sum_{j=1}^n |a_j^\top x - b_j|$. Is f convex ? differentiable ? separable ?
2. For $z \in \mathbb{R}$, calculate $\phi(z) = \max_{u \in [-1,1]} uz$.
3. Express $\max_{u \in [-1,1]^m} \langle u, Ax - b \rangle$ as a function of x.
4. Let $u(x)$ be any element of $\arg\max_{u \in [-1,1]^m} \langle u, Ax - b \rangle$.
   Show that this $\arg\max$ is never empty and that $f(x) = \langle u(x), Ax - b \rangle$.
5. Show that $A^\top u(x) \in \partial f(x)$.
6. Show that $\|A^\top u(x)\|_2 \leq \sqrt{m}\, \|A\|$ where $\|A\|$ is the operator norm of A.
7. Using the results of Exercise 2, show that f is Lipschitz continuous and give a bound on $f(\bar{x}_k) - f(x_*)$ for the choice $\gamma_k = \frac{1}{\sqrt{k+1}}$.